Science-2

TUTORIAL ASSIGNMENT 1

Aritra Banerjee Roll No: 201812006

1. DotPlot Analysis

(a) Dottup --- k-tuple search

(b) Dotmatcher – sliding window

Run these programs for different k-tuples (Dottup) and different window size and threshold (Dotmatcher).

Given are the sequences of spike glycoprotein (both DNA and protein) of the following: (1) SARS-CoV (2003), MERS-COV (2012), and (3) SARS-CoV2 (2019). Submit the results of Dottup and Dotmatcher and answer the following Qs:

1. Identify SARS-CoV2 is similar to which of the earlier two viruses
2. Is it easy to identify the similarity using DNA or protein sequences? Give reasons
3. Submit the graphs and give the k-tuple values used, and window size and threshold values used

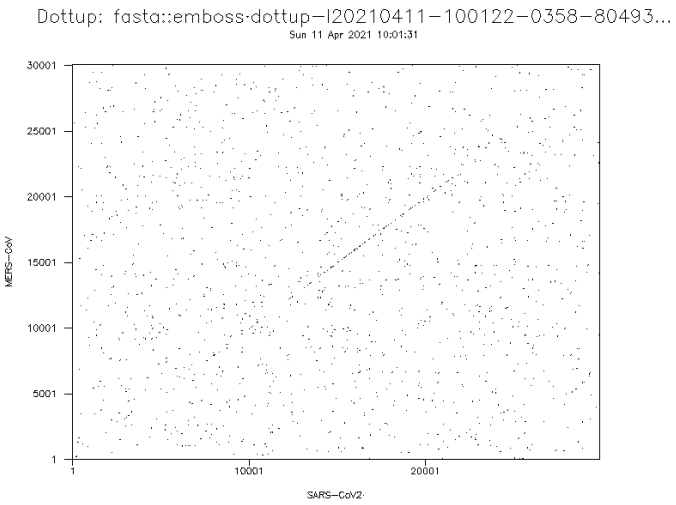
Solution:

1. SARS-COV2(2019) is most similar to SARS COV (2003). We infer this from the DotPlot analysis graphs as drawn below.
2. It is more feasible to use Protein because the graphs for DNA matches for SARS-CoV and MERS-CoV with SARS-CoV2 are noisier compared to their Protein match counterparts. This is true for the plots generated by both Dottup and Dotmatcher.
3. Parameters: Word size = 10 (for Dottup)

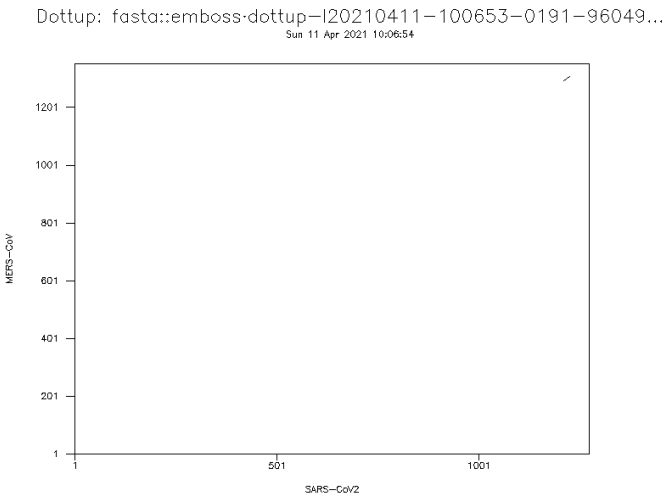
Parameters: Word size = 10, Threshold = 50, Matrix = BLOSUM62/DNAfull (for Dotmatcher)

DNA and Protein matches for MERS-CoV and SARS-CoV2 using Dottup :--

DNA match:



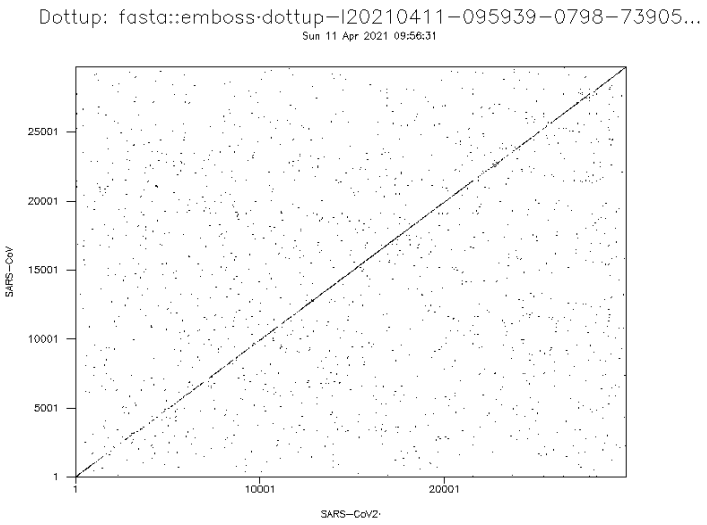
Protein Match:



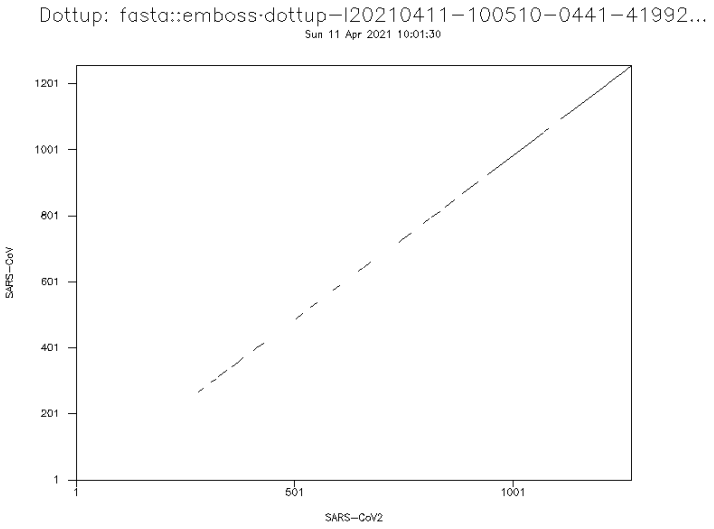
From the above two images, we can infer that there are very few to almost no matches(no line of any kind can be seen clearly) between the two types of coronavirus. So, **they are not very similar**.

DNA and Protein matches for SARS-CoV and SARS-CoV2 using Dottup:--

DNA match:



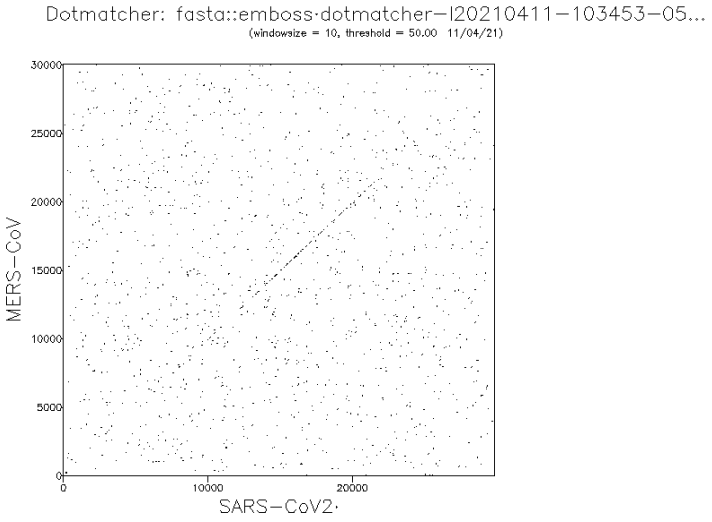
Protein match:



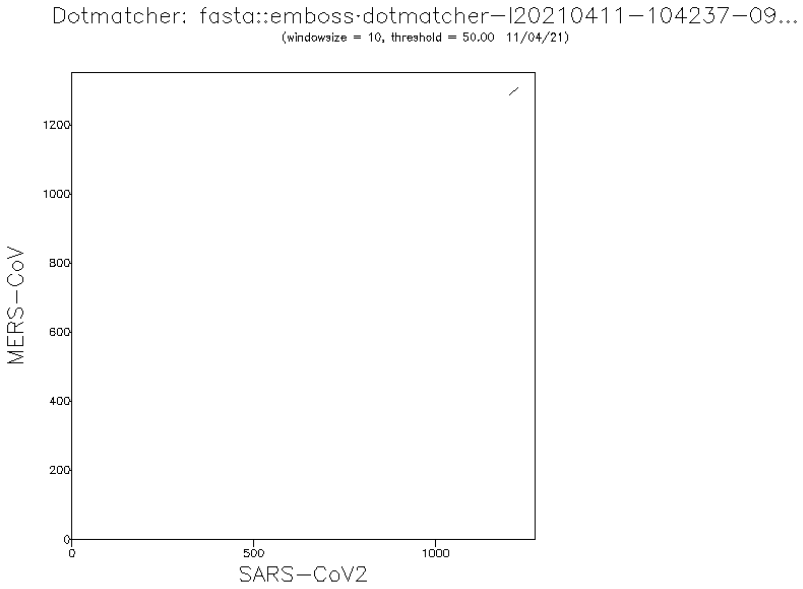
From these two images we observe straight line(for DNA match) and almost straight line for (Protein match). So, we can say that **the two types of coronavirus are quite similar**.

DNA and Protein matches for MERS-CoV and SARS-CoV2 using Dotmatcher :--

DNA match:



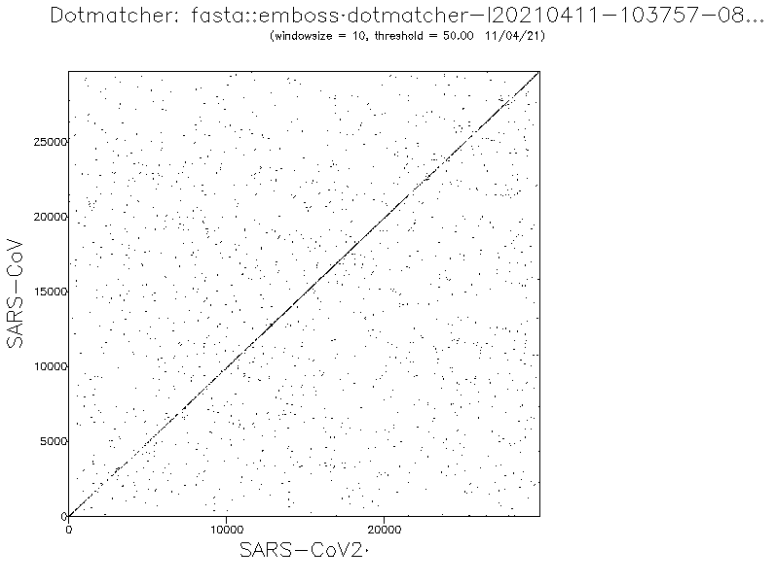
Protein match:



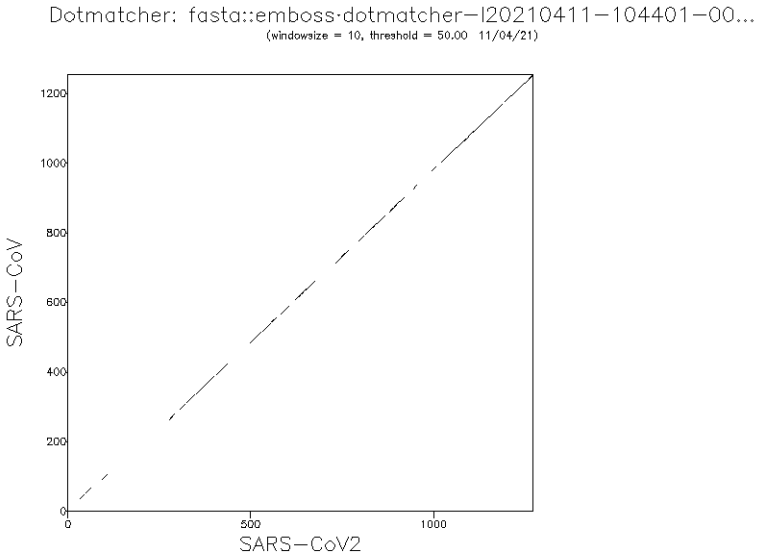
These two graphs generated by the Dotmatcher are very similar to that for Dottup

DNA and Protein matches for SARS-CoV and SARS-CoV2 using Dotmatcher :--

DNA match:



Protein match:



The two graphs generated by Dotmatcher are very close to that by Dottup

2(A). Pairwise Alignment

Perform pairwise alignment of spike glycoprotein of SARS-CoV2 with that of SARS-CoV, both at the DNA and protein level, using programs ‘needle’ and ‘water’. Answer the following Qs.

(1) What is percentage identity and percentage similarity at DNA level and protein level? Which is larger and why, give reasons.
(2) What is the difference between identity and similarity?
(3) Is there any difference in the global and local alignments of these two sequences?
(4) Submit the alignment giving the scoring scheme and gap penalties used.

Solution:

(1) Percentage identity and similarities are given below:

According to needle (checks for global alignments):-

- At DNA level,
    - Identity: 2833/3859 (85.5%)
    - Similarity: 2833/3859 (85.5%)
- At the Protein level,
    - Identity: 918/1308 (76.3%)
    - Similarity: 1057/1308 (86.9%)

According to water (checks for local alignments):-

- At the DNA level,
    - Identity: 2880/3933 (85.6%)
    - Similarity: 2880/3933 (85.6%)
- At the Protein level,
    - Identity: 908/1213 (76.3%)
    - Similarity: 1043/1213 (86.9%)

From the above we understand that the DNA levels for both identity and similarity are the same. However, at Protein level, similarity is much larger than identity. This is because protein sequences have common ancestry and as a result, similar properties. The actual elements in the sequence might be different which gives a lower identity percentage.

(2) Sequence identity is the number of characters that match exactly between two different sequences. Gaps are not considered here. Similarity is a measure of how much the two sequences resemble each other. It depicts the length to which the residues are aligned. Similar sequences have similar properties.
(3) At DNA level, we can't see any difference between local and global alignments for the two sequences. But, in the protein level, local similarity is higher (86%) compared to global similarity (80.6%).
(4) Please see the following files:
    (a) 2A_DNA_global.txt
    (b) 2A_Protein_global.txt
    (c) 2A_DNA_local.txt
    (d) 2A_Protein_local.txt

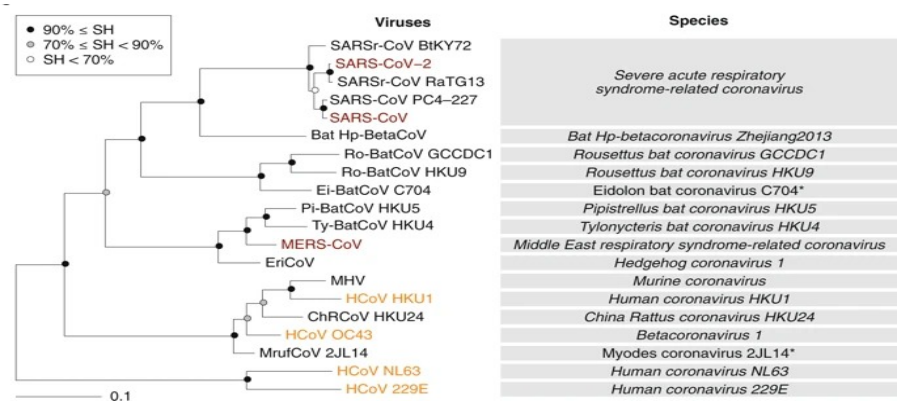
2(B). Pairwise Alignment

Perform pairwise alignment of spike glycoprotein of SARS-CoV2 with that of SARS-CoV, both at the DNA and protein level, using programs 'needle' and 'water'. Answer the following Qs.

(1) Based on the sequence alignment, can you say that the two proteins are homologs, i.e., related?
(2) 2. Are you able to make this inference from alignment of DNA sequences or protein sequences?

Solution:

(1) Two proteins are homologous if they have a common ancestor (as shown below) and so their proteins are homologs.



(2) Similarity may or maynot imply homology and vice-versa as shown below:

According to needle (checks for global alignments):-

- At DNA level,
    - Identity: 2833/3859 (57.6%)
    - Similarity: 2833/3859 (57.6%)
- At the Protein level,
    - Identity: 918/1308 (30.1%)
    - Similarity: 1057/1308 (45.8%)

According to water (checks for local alignments):-

- At the DNA level,
    - Identity: 2880/3933 (57.8%)
    - Similarity: 2880/3933 (57.8%)
- At the Protein level,
    - Identity: 908/1213 (30.4%)
    - Similarity: 1043/1213 (46.0%)

According to the rule of thumb, two sequences are approximately homologous if sequence similarity is above ~40%. Two sequences are homologous if their identity is more than 30% over entire lengths. Thus, we can say that MERS-CoV and SARS-CoV are homologous.

The files used are:

(a) 2B_DNA_global.txt
(b) 2B_Protein_global.txt
(c) 2B_DNA_local.txt
(d) 2B_Protein_local.txt

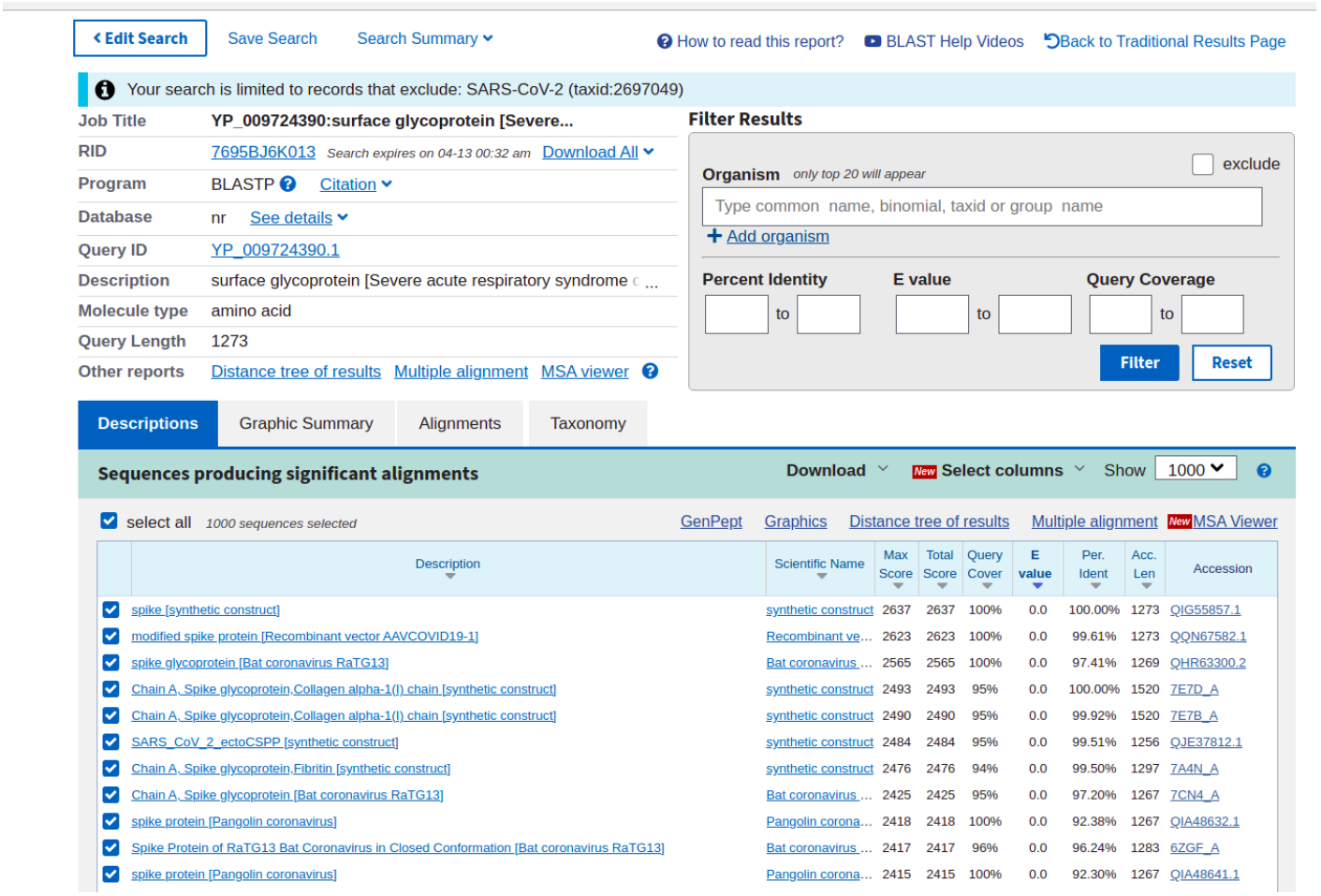
3. Database Search [BLASTp]

Perform protein database search using spike glycoprotein of SARS CoV2 (Acc. ID: YP_009724390.1) as query and answer the following Qs:

1. Which is the closest homolog of the query sequence?
2. Give the score, percentage identity, percentage similarity, length of the alignment, and the expect or e-value.
3. Do you find the spike glycoprotein of SARS-CoV as one of the hits? Does the percentage identity and percentage similarity results match with the alignment obtained using 'water'? What is the significance of this alignment?
4. It was speculated that SARS-CoV2 has come from bat. Do you find any relation of spike glycoprotein of SARS-CoV2 with that of bat SARS coronavirus spike glycoprotein? What is identity, similarity, length of alignment, score and e-value?

Solution:

On conducting the database search, we get this:



1. The closest homolog to the SARS-CoV2 spike glycoprotein seems to be the **spike glycoprotein [Bat coronavirus RaTG13]**.
2. The details required:
    a. Score: 2565
    b. Percentage identity: 97%
    c. Percentage similarity: 98%
    d. Length of alignment: 1269
    e. E-value: 0.0

3. Yes, we do find spike glycoprotein of SARS-CoV as one of the hits. Yes, the values of percentage identity (76%) and similarity (87%) match with the corresponding results obtained from 'water'. The significance of this alignment is that SARS-CoV2 is the nearest natural variant of SARS-CoV which infects humans.
4. Yes, Bat Coronavirus is found to be the closest homolog of this sequence. Based on the results we can conclude that there are a lot of similarities between the spike glycoproteins of these two viruses. The parameters are:
    a. Score: 2565
    b. Percentage identity: 97%
    c. Percentage similarity: 98%
    d. Length of alignment: 1269
    e. E-value: 0.0

4. Database Search [BLASTn]

To identify from which species it might have jumped on to humans, perform BLAST database search using its genomic sequence (Acc. ID: NC_045512.2). [Hint: Set the number of target sequences to 1000 or more, and pay attention to query sequence coverage before reporting your results].

(a) Which BLAST nucleotide program would you use and why for genomic sequence comparisons?

(b) What do the top hits correspond to? Take any two top hits and analyze their alignment with the query and report the variation.

(c) List top 5 non-SARS-COV-2 sequence hits from different species and give the sequence coverage, %age identity, and e-value. Which coronavirus species is closest to

SARS-COV-2? What do you infer from this result? Did SARS-COV-2 come from bat or pangolin, as earlier expected?

Solution:

(a) We should use the **blastn** nucleotide program for genomic sequence comparison. This is because it helps us to check for sequences which may not have a very high similarity. This helps us to understand the evolution of these viruses and find distant homologs of viruses.

(b) Top hits correspond to synthetic spike glycoproteins. Two top hits are Bat coronavirus RaTG13 and Pangolin coronavirus isolate MP789.

- (i) For bat coronavirus:
  - (1) Length of alignment: 29885
  - (2) Percentage identity: 96%
  - (3) Score: 48667 / 53973
- (ii) For pangolin coronavirus:
  - (1) Length of alignment: 29521
  - (2) Percentage identity: 90%
  - (3) Score: 40124 / 44498

(c) Top 5 non-SARS-COV-2 hits from different species are:

- (i) Bat coronavirus RaTG13:
  - (1) Sequence coverage: 29855
  - (2) Percentage identity: 96%
  - (3) E-value: 0.0
- (ii) Pangolin coronavirus isolate MP789:
  - (1) Sequence coverage: 29521
  - (2) Percentage identity: 90%
  - (3) E-value: 0.0
- (iii) Bat coronavirus RacCS203:
  - (1) Sequence coverage: 29832

(2) Percentage identity: 94%

(3) E-value: 0.0

(iv) Bat SARS-like coronavirus isolate bat-SL-CoVZC45:

(1) Sequence coverage: 29802

(2) Percentage identity: 88%

(3) E-value: 0.0

(v) Rhinolophus affinis coronavirus isolate LYRa11:

(1) Sequence coverage: 29805

(2) Percentage identity: 81%

(3) E-value: 0.0

Bat coronavirus RaTG13 is the closest to SARS-CoV2. From the result above we infer that SARS-CoV2 might have evolved from these coronaviruses owing to their high identity percentage and alignment. From the results, we can confirm the statement that SARS-CoV2 has come from the bat or pangolin.

5. UniProt and GenBank

Find out the size of protein database, UniProt, and nucleotide database, GenBank. Compute No. of matrix cells to be computed using DP for:

1. Performing search in protein database, UniProt, and nucleotide database, GenBank and the time required assuming query sequence of length 1000 bases.
2. Comparing Human Chr 1 ~249Mbp with a query sequence of 1000 bases using DP, and comparing it with Chr 1 of Mouse (~195Mbp)? What is the memory or space requirement in the two cases?

Solution:

Size of uniplot = 177754527 sequences = 59974041839 amino acids (Source:https://www.ebi.ac.uk/uniprot/TrEMBLstats)

Size of GenBank = 216214215 sequences = 399376854872 bases (Source: https://www.ncbi.nlm.nih.gov/genbank/statistics/)

Memory complexity required for checking the similarity of two sequences = **m*n**

Where **m** and **n** are the number of bases/acids in the two sequences

Time complexity required for the same = **O(mn)**

1. Length of query sequence = 1000

   Total number of matrix cells that would be made for comparing DNA sequence

   = total number of bases in GenBank * query sequence

   = 399376854872 * 1000

   = **399376854872000**

   Total number of matrix cells that would be made for comparing protein sequence

   = total number of acids in uniplot * query sequence

   = 59974041839 * 1000

   = **59974041839000**

   Number of iterations that would be made for comparing DNA sequence

   = total number of bases in GenBank * query sequence

   = 399376854872 * 1000

= **399376854872000**

If we assume that 10^7 iterations would be completed in 1 second, then,

Total time taken = **39937685.4872**

Number of iterations that would be made for comparing protein sequence

= total number of acids in uniplot * query sequence

= 59974041839 * 1000

= **59974041839000**

If we assume that 10^7 iterations would be completed in 1 second, then,

Total time taken = **5997404.1839**

2. Human Chr 1

    m = 1000

    n = 249Mbp = 249 * 10^6 * 2 = 498 * 10^6

    Memory complexity = m * n = 498 * 10^9

    Time complexity = m * n = 498 * 10^9

Mouse Chr 1

    m = 1000

    n = 195Mbp = 195 * 10^6 * 2 = 390 * 10^6

    Memory complexity = m * n = 390 * 10^9

    Time complexity = m * n = 390 * 10^9